

# A sparse representation for potential energy surface

Atsuto Seko,<sup>1,2,\*</sup> Akira Takahashi,<sup>1</sup> and Isao Tanaka<sup>1,2,3</sup>

<sup>1</sup>*Department of Materials Science and Engineering, Kyoto University, Kyoto 606-8501, Japan*

<sup>2</sup>*Center for Elements Strategy Initiative for Structure Materials (ESISM), Kyoto University, Kyoto 606-8501, Japan*

<sup>3</sup>*Nanostructures Research Laboratory, Japan Fine Ceramics Center, Nagoya 456-8587, Japan*

(Dated: April 1, 2014)

We propose a simple scheme to estimate potential energy surface (PES) with which the accuracy can be easily controlled and improved up to the level of the density functional theory (DFT) calculations. It is based on a model selection within the framework of linear regression using the least absolute shrinkage and selection operator (LASSO) technique. Basis functions are selected from a systematic large set of candidate functions. The sparsity of PES significantly reduces the computational demands for evaluation of the energy and the force in molecular dynamics simulations without losing the accuracy. The usefulness of the scheme is well demonstrated for describing elemental metals of Na and Mg.

PACS numbers: 31.50.Bc, 34.20.-b, 65.40.-b, 71.15.Pd

About sixty years ago, molecular dynamics (MD) was proposed as a tool to model a collection of interacting atoms within classical mechanics[1]. Accuracy of potential energy surface (PES) with respect to the atomic positions is crucial in performing MD simulations. PES determines forces acting on atoms that originate from atomic interactions and thereby motions of atoms. A reliable PES can be obtained by computing the energy and forces acting on atoms for an atomic configuration at each time step directly from density functional theory (DFT) calculations[2, 3], which is called the first-principles MD calculations[4]. Since this approach is computationally demanding, many different kinds of interatomic potentials have been used instead; they include Lennard-Jones, Morse, embedded atom method (EAM) and Tersoff potentials[5–10]. Parameters in the potential functions were optimized by fitting to a set of experimental data or first principles results. Although such interatomic potentials have been used widely, its accuracy has been validated only for limited cases.

Recently, an alternative approach for estimating PES from a large set of DFT energies was demonstrated[11, 12]. In these techniques, the information of atomic positions in a structure was transformed into some descriptors. Then, nonlinear regression techniques such as the neural network[11] and Gaussian process[12] were used to bridge the energy and descriptors because a complex function of PES with respect to the descriptors is required to be estimated without a priori knowledge of the target material. Using these techniques, accuracy of PES is generally much better than that of conventional interatomic potentials. Another advantage is their applicability to a wide range of materials including metallic[13, 14], covalent[11, 12, 15] and ionic materials[16]. In these frameworks, however, the accuracy of the PES strongly depends on the selection of descriptors and regression techniques. Although the use of some descriptors has already been proposed[17–19], there is no clear answer

which descriptors should be used for the target material. In addition, since nonlinear regression techniques generally require many parameters to be optimized, they often cause low predictive power for structures that are not included in the input data set. We need a simple procedure for estimating PES with which the accuracy of the PES can be easily controlled and improved.

In this Letter, we propose to use a simple regression technique and a systematic set of simple basis functions as descriptors to estimate the PES from a large set of DFT calculations. The procedure makes it easy to control the accuracy of the PES. This is included in the framework of the linear regression although the PES is a complex function with respect to the atomic positions. Compared to nonlinear regression techniques, the linear regression has a number of advantages as follows: 1) Accuracy can be controlled in a transparent manner. 2) Regression coefficients are generally determined quickly using a standard least-square technique. 3) The number of the regression coefficients does not explicitly depend on the size of the input data-set. We demonstrate the use of two methods on the estimation of PES for elemental metals of Na and Mg. One is the linear ridge regression technique[20] using all basis functions in the given set of basis functions. The other uses the least absolute shrinkage and selection operator (LASSO) technique[21] in order to optimize the selection of basis functions. The LASSO enables us to obtain a well-optimized sparse expression for the PES with a small number of non-zero coefficients.

A linear model for describing the energy for a structure is shown in Fig. 1. This model is invariant for translation and exchange of atoms and stands on a widely-accepted idea that the total energy for a structure is equal to the sum of atomic energies[11, 12]. The total energy of the structure  $i$  is expressed as  $E^{(i)} = \sum_j E^{(i,j)}$ , where  $E^{(i,j)}$  denotes the contribution of the atom  $j$  to the total energy of the structure  $i$ . Then, a linear relationship between the

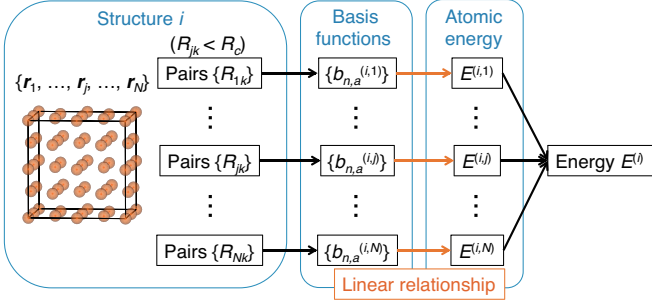


FIG. 1. Relationship between atomic positions and total energy in the linear model for the PES of structure  $i$ . It is assumed that the energy of atom  $j$ ,  $E^{(i,j)}$ , and the set of basis functions for atom  $j$ ,  $\{b_{n,a}^{(i,j)}\}$ , has a linear relationship. The set of basis functions is here calculated from the set of pair distances  $\{R_{jk}^{(i)}\}$  smaller than the cutoff radius  $R_c$ , which are obtained from positions of all atoms.

atomic energy and  $M$  basis functions is introduced as

$$E^{(i,j)} = \mathbf{w}^\top \mathbf{b}^{(i,j)}, \quad (1)$$

where  $\mathbf{w} = [w_1, \dots, w_M]^\top$  and  $\mathbf{b}^{(i,j)} = [b_1^{(i,j)}, \dots, b_M^{(i,j)}]^\top$  denote the regression coefficients and basis functions for the atom  $j$  of the structure  $i$ , respectively. By applying the same regression coefficients to identical atomic species, the total energy is derived as

$$E^{(i)} = \mathbf{w}^\top \mathbf{c}^{(i)}, \quad (2)$$

where  $M$ -vector  $\mathbf{c}^{(i)}$  satisfies the equation of  $\mathbf{c}^{(i)} = \sum_j \mathbf{b}^{(i,j)}$ .

Here, a simple form of functions is implemented as basis functions. The basis function, which is the power  $a$  of  $n$ -th element,  $b_{n,a}^{(i,j)}$ , is written as

$$b_{n,a}^{(i,j)} = \left[ \sum_k f_n(R_{jk}^{(i)}) \cdot f_c(R_{jk}^{(i)}) \right]^a, \quad (3)$$

where  $a$  is a positive integer and  $R_{jk}^{(i)}$  denotes the distance between atoms  $j$  and  $k$  of structure  $i$ . The sum is taken over all atoms within a cutoff radius  $R_c$ . For  $f_n(R_{jk}^{(i)})$ , we adopted various kinds of systematic and analytical pairwise functions, which are Gaussian, cosine, Bessel, Neumann, polynomial and Gaussian-type orbital (GTO) functions.  $f_c(R_{jk}^{(i)})$  is a smooth pairwise cutoff function which is exactly zero at a distance greater than the cutoff radius  $R_c$ . We use a cosine-based cutoff function as used in Ref. 11. Since a product of  $f_n(R_{jk}^{(i)})$  and  $f_c(R_{jk}^{(i)})$  is pairwise, an exponential form of the sum of the pairwise functions is introduced in order to take many body effects into account. It should be noted that the use of the pairwise potential causes well-known serious problems for the description of PES[9, 10]. For instance, pairwise interatomic potentials satisfies the Cauchy-relationship that

elastic constants have the relationship of  $C_{12} = C_{44}$  in fcc crystals; this is an artifact. The cohesive energy is forced to exhibit a linear dependence on the coordination number.

To begin with, the PESs for elemental metals of Na and Mg described by the linear model were estimated from a large set of DFT calculations using the linear regression techniques. As training data for the regressions, 1600 configurations were prepared on the basis of face-centered-cubic (fcc), body-centered-cubic (bcc), hexagonal-closed-packed (hcp) and simple cubic (sc) structures for both Na and Mg. They were generated by random distortions on the ideal fcc, bcc, hcp and sc structures, in which the atomic positions and lattice constants were fully optimized. In addition to the training data, 400 configurations were prepared by the same procedure as the training data in order to examine the predictive power for structures that were not included in the training data. In total 2000 configurations for both Na and Mg, DFT calculations were performed using the plane-wave basis projector augmented-wave (PAW) method[22, 23] within the generalized gradient approximation (GGA)[24] as implemented in the VASP code[25, 26]. The total energies converged to less than  $10^{-2}$  meV/supercell. For the ideal structures, the atomic positions and lattice constants were optimized until the residual forces became less than  $10^{-3}$  eV/Å.

Subsequently, we constructed PESs for Na and Mg from the training data using the linear ridge regression[20]. The linear ridge regression is one of shrinkage methods and shrinks the regression coefficients by imposing a penalty. When matrix  $\mathbf{X} = [\mathbf{c}^{(1)}, \dots, \mathbf{c}^{(N)}]^\top$  is composed of basis functions for the training data, the ridge coefficients minimize a penalized residual sum of squares expressed as

$$\|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 + \lambda \|\mathbf{w}\|_2^2, \quad (4)$$

where  $\mathbf{y}$  denotes DFT energies of the training data and  $\|\cdot\|_2$  means  $L_2$ -norm. This is referred to as  $L_2$  regularization. The regularization coefficient  $\lambda$  controls the magnitude of the penalty. The solution is given only by matrix operations as  $\mathbf{w} = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y}$ , where  $\mathbf{I}$  denotes the unit matrix. When regression coefficients of many correlated variables in a linear regression model are determined without including the penalty term, they can be poorly determined and exhibit large variance. The penalty on the coefficients alleviates this problem.

For each PES constructed using a set of basis functions, we calculated the root-mean-square error (RMSE) between energies for the test data predicted by the DFT calculation and those by the PES; this can be regarded as the predictive power of the PES. Table I shows the RMSEs of PESs for Na and Mg built from 240 cosine-type systematic basis functions, where the RMSE converges with respect to the number of basis functions. PESs constructed from basis functions of  $a = 1$  have low predictive

TABLE I. RMSEs for the test data of PESs obtained by the linear ridge regression and 240 cosine-type basis functions,  $f_n(R) = \cos(\beta R)$ , with  $R_c = 7.0$  and  $\lambda = 0.001$ . The interval and minimum value of  $\beta$  are fixed to 0.05, and the maximum values of  $\beta$  and  $a$  controls the number of basis functions.

	Na (meV/atom)	Mg (meV/atom)
Cosine ( $a_{\max} = 1$ )	7.3	11.8
Cosine ( $a_{\max} = 2$ )	1.6	2.6
Cosine ( $a_{\max} = 3$ )	1.4	1.6

powers for both Na and Mg, in which only pairwise interactions are considered. On the other hand, the increase of the maximum value of  $a$ ,  $a_{\max}$ , improves the predictive powers of PESs substantially. Using 240 cosine-type basis functions with  $a_{\max} = 3$  and  $R_c = 7.0$  Å, the RMSEs for Na and Mg are 1.4 and 1.6 meV/atom, respectively. By increasing the cutoff radius to  $R_c = 9.0$  Å, the RMSE reaches to a considerably-small value of 0.4 meV/atom for Na, whereas it remains almost unchanged for Mg.

Then, we employed single other types of systematic basis functions with  $a_{\max} = 3$ , i.e. Gaussian-, Bessel- and Neumann-type basis functions in addition to cosine-type basis functions. Figures 2 (a) and (b) show the dependence of the RMSE on the number of the basis functions for Na and Mg, respectively. When a small number of basis functions of  $< 20$  are used, the Neumann basis-set is the best one among the four basis-sets for both Na and Mg, although the RMSE is rather large. On the other hand, when the number of basis functions is increased, the cosine basis-set is satisfactory for increasing the accuracy for both Na and Mg although plenty of basis functions are needed for the convergence of the RMSE for Mg.

After applying the single type of basis functions to the estimation of PES, we considered all combinations ( $= 2^4 - 1 = 15$ ) of the four basis-sets with  $a_{\max} = 3$ . Gaussian and cosine basis-sets are composed of 120 basis functions. Bessel and Neumann basis-sets are composed of 60 and 30 basis functions, respectively. As a result, any combinations do not improve the RMSE for Na, whereas the combination of Gaussian, cosine and Bessel basis-sets gives the best prediction for Mg, whose RMSE is 0.9 meV/atom.

These results indicate that the energy can be expressed by a linear relationship with simple basis functions depending only on distances between atoms. However, it appears that this method leads to use a large number of unnecessary basis functions for describing the PES. In order to avoid the problem, the LASSO technique[20, 21] is applied, which enables us not only to provide a solution for a linear regression but also to obtain a sparse representation with a small number of non-zero regression coefficients. The LASSO is another shrinkage method, similar to the ridge regression. Using a large set of can-

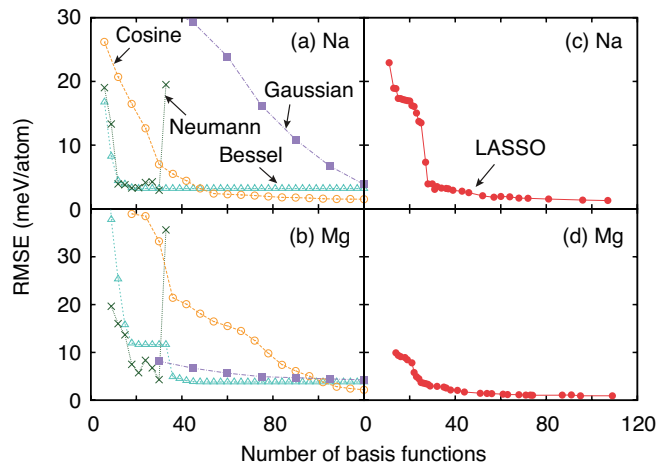


FIG. 2. RMSEs for the test data of the PESs constructed by the linear ridge regression using various type of basis functions with  $a_{\max} = 3$ ,  $R_c = 7.0$  and  $\lambda = 0.001$  (a) for Na and (b) for Mg. RMSEs optimized by the LASSO are also shown (c) for Na and (d) for Mg. For Bessel and Neumann basis-sets, the number of basis functions is controlled by the maximum order of the basis functions. For Gaussian basis-set, the number of basis functions is controlled by the maximum values of internal parameters of Gaussian.

didates composed of various kinds of systematic basis functions for the LASSO, three kinds of unknown features can be simultaneously optimized, i.e. the type of basis functions, internal parameters of the basis functions and the number of basis functions.

The LASSO minimization function is defined as

$$\|X\mathbf{w} - \mathbf{y}\|_2^2 + \lambda\|\mathbf{w}\|_1, \quad (5)$$

where  $\|\cdot\|_1$  denotes  $L_1$ -norm. The  $L_2$  ridge penalty in Eqn. (4) is replaced by the  $L_1$  LASSO penalty. The parameter  $\lambda$  controls the trade-off relationship between the sparsity and accuracy. The LASSO solution is computed using a general quadratic programming technique, where efficient algorithms are available.

Candidates of basis functions were composed of a great deal of Gaussian-, cosine-, Bessel-, Neumann, polynomial-, GTO-type basis functions, generated with fine intervals for internal parameters of the basis functions. The total number of candidate basis functions was 8455, which was much larger than the number of structures in the training data. Sparse representations were then extracted from the set of candidate basis functions by the LASSO. Figures 2 (c) and (d) show the RMSEs of PESs obtained by the LASSO for Na and Mg, respectively. The RMSE of the LASSO PES decreases more rapidly than those of PESs constructed from single type of basis functions. In other words, the LASSO PES requires much smaller number of basis functions than the ridge regression. For Na, a sparse representation with the

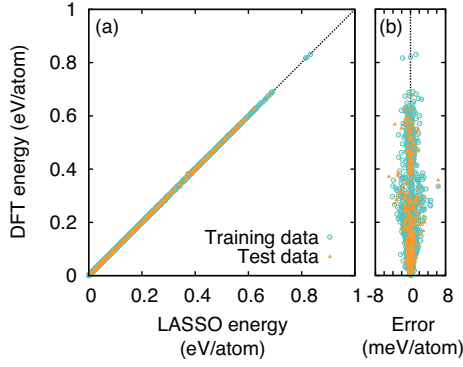


FIG. 3. (a) Energies predicted by the LASSO PES and DFT for Mg, measured from the energy of hcp ideal structure. (b) Differences between energies predicted by the LASSO PES and DFT energy.

RMSE of 1.3 meV/atom was obtained only using 107 basis functions, which has almost the same accuracy as the PES constructed by 240 cosine-type basis functions with the RMSE of 1.4 meV/atom. Moreover, the LASSO PES is composed of about eighteen times smaller number of regression coefficients than a neural network potential with 1901 coefficients[13], although the RMSE of the LASSO PES is slightly larger than that of the neural network potential of 0.91 meV/atom.

It is apparent that the LASSO is more advantageous for Mg than for Na. The obtained sparse representation with 95 basis functions for Mg shows the RMSE of 0.9 meV/atom, which is almost a half of the RMSE in the PES constructed from 240 cosine-type basis functions. Figure 3 shows the comparison of energies predicted by the LASSO PES and DFT in Mg. As can be seen in Fig. 3, there is little difference between the prediction errors for training and test data. In addition, the dependence of the prediction error on the energy is not clearly observed despite a wide range of structures are included in both training and test data.

Once the PES is constructed within the linear model, forces acting on atoms and stress can be analytically computed using Eqns. (2) and (3). Here the accuracy of the LASSO PES for the force calculation can be examined by comparing phonon dispersions computed by the LASSO PES and DFT. The phonon dispersions and related thermodynamic properties were calculated by the supercell approach[27]. The phonon calculations were performed using PHONOPY code[28]. Figures 4 (a) and (b) show phonon dispersions and specific heats at constant volume for hcp Mg, respectively, using the LASSO PES and DFT calculation. Since the phonon dispersion of the LASSO PES does not differ from that of the DFT, the specific heats agree well to each other. This demonstrates that the LASSO PES is accurate enough to perform atomistic simulations in solids with the similar accuracy to

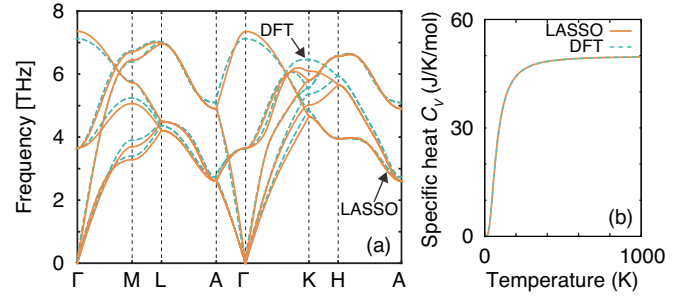


FIG. 4. (a) Phonon dispersion and (b) specific heat at constant volume for hcp Mg obtained from the LASSO PES (solid lines), compared with those obtained by the DFT calculation (dotted lines). For evaluating the dynamical matrix, each atomic position is displaced by 0.01 Å.

the DFT calculation.

In conclusion, we have introduced a simple scheme to estimate PES with which the accuracy can be easily controlled and improved up to the level of the DFT calculations. We have applied it to describe elemental metals of Na and Mg. We found that the energy can be expressed by a linear relationship with simple basis functions depending only on distances between atoms. Using LASSO, a sparse set of meaningful basis functions for expressing PES can be easily extracted from a great deal of candidate functions. As the result, we have obtained sparse PES with the prediction errors of 1.3 and 0.9 meV/atom for Na and Mg, respectively. The present method can accelerate to increase the accuracy of atomistic simulations with decreasing the computational demands.

This study is supported by a Grant-in-Aid for Scientific Research (A) and a Grant-in-Aid for Scientific Research on Innovative Areas "Nano Informatics" (grant number 25106005) from Japan Society for the Promotion of Science (JSPS).

\* seko@cms.mtl.kyoto-u.ac.jp

- [1] B. Alder and T. Wainwright, J. Chem. Phys. **31**, 459 (1959).
- [2] P. Hohenberg and W. Kohn, Phys. Rev. **136**, B864 (1964).
- [3] W. Kohn and L. J. Sham, Phys. Rev. **140**, A1133 (1965).
- [4] R. Car and M. Parrinello, Phys. Rev. Lett. **55**, 2471 (1985).
- [5] J. Lennard-Jones, Proc. R. Soc. London, Ser. A **106**, 463 (1924).
- [6] P. M. Morse, Phys. Rev. **34**, 57 (1929).
- [7] J. Tersoff, Phys. Rev. Lett. **56**, 632 (1986).
- [8] I. M. Torrens, *Interatomic potentials* (Academic Press, New York, 1972).
- [9] A. Carlsson, Solid State Physics **43**, 1 (1990).
- [10] M. Finnis, *Interatomic forces in condensed matter* (Oxford University Press, New York, 2003).

- [11] J. Behler and M. Parrinello, Phys. Rev. Lett. **98**, 146401 (2007).
- [12] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, Phys. Rev. Lett. **104**, 136403 (2010).
- [13] H. Eshet, R. Z. Khaliullin, T. D. Kühne, J. Behler, and M. Parrinello, Phys. Rev. B **81**, 184107 (2010).
- [14] N. Artrith and J. Behler, Phys. Rev. B **85**, 045439 (2012).
- [15] G. C. Sosso, G. Miceli, S. Caravati, J. Behler, and M. Bernasconi, Phys. Rev. B **85**, 174103 (2012).
- [16] N. Artrith, T. Morawietz, and J. Behler, Phys. Rev. B **83**, 153101 (2011).
- [17] J. Behler, J. Chem. Phys. **134**, 074106 (2011).
- [18] K. J. Jose, N. Artrith, and J. Behler, J. Chem. Phys. **136**, 194111 (2012).
- [19] A. P. Bartók, R. Kondor, and G. Csányi, **87**, 184115 (2013).
- [20] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. (Springer, New York, 2009).
- [21] R. Tibshirani, J. Roy. Stat. Soc. B , 267 (1996).
- [22] P. E. Blöchl, Phys. Rev. B **50**, 17953 (1994).
- [23] G. Kresse and D. Joubert, Phys. Rev. B **59**, 1758 (1999).
- [24] J. P. Perdew, K. Burke, and M. Ernzerhof, Phys. Rev. Lett. **77**, 3865 (1996).
- [25] G. Kresse and J. Hafner, Phys. Rev. B **47**, 558 (1993).
- [26] G. Kresse and J. Furthmüller, Phys. Rev. B **54**, 11169 (1996).
- [27] K. Parlinski, Z. Q. Li, and Y. Kawazoe, Phys. Rev. Lett. **78**, 4063 (1997).
- [28] A. Togo, F. Oba, and I. Tanaka, Phys. Rev. B **78**, 134106 (2008).